REGRESSION OUTPUT EXPLAINED

Example → The winter olympics : Does a country's latitude affect their medal tally? Total countries participate are 25.

Y Variable → Number of medals X → i) Latitude ii) Average elevation iii) Log population.

Inference & Significance

Suppose, Medals$_i$ = $\beta_0 + \beta_1$ (Latitudes), it defines the relationship between Medals and Latitudes.

If β_1 is positive, both latitudes & medal goes up. ie, β_1 is actual affect.

If β_1 is negative, latitudes goes up and medal goes down.

If β_1 is zero, there will be no affect of latitudes on medal.

Normally, we try to find β_1 as non-zero value, so that it should have any kind of relationship between Independent and dependent variable.

So for Winter Olympics, can we infer a relationship between

number of medals$_i$ = $\beta_0 + \beta_1$ (latitude) + β_2 (elevation) + β_3 (log population)

First section – ANNOVA Section / Analysis of Variance

How much variation is there in the dependent variable?

Total medals = 33, 28, 26, 25, . . . , 3, 1, 1, 1

Top country have 33 medals   Lowest medal by a country

Average (Total medals) = 11.3 medals

SS, sum of square, how spread out our data is. It should be decent low.

$$SS = \Sigma(x_i - \bar{x})^2 = (33 - 11.3)^2 + (28 - 11.3)^2 + \dots = 1393.76$$

So, 1393.76 is amount of variation in the y variable.

So, through Independent Variable (X variables) we will try to explain the 1393.76.

Annova output →

Source	SS	df	MS
Model	439.2	3 (K)	146.4
Residual	954.4	21 = (n-K-1)	45.4
Total	1393.7	24 (n-1)	58.07

df → degree of freedom

df = 3 means we are using 3 Independent variable

24 is countries participate -1. 25-1=24

21 = n-K-1 => 25-3-1 = 21.

→ Model is explaining 439.4 out of 1393.7.

→ Residual / Error is 954.4 out of 1393.7

How much "explaining" is model doing? → $R^2 = 439.2/1393.7 = 0.315$, so 31% is the variation in y explained by X variables.

So, 69% is still remaining / not explained.

Is this model with 3 explanatory variables better than a model with 0 explanatory variables?

↳ $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

MS is mean square, SS divide by df.

Calculate F statistic (Refer to MS) = 146.4/45.4 = 3.22, so 3.22 > 0.05 therefore reject H0, at 5% level of significance

45.4 is some time called Standard Error or Standard error Residual (SER) or Mean Square Error (MSE). Higher the MSE, bad is the model.

Another Output (Continued from last Output) Automatically calculated.

Number of observation	= 25
F(3, 21)	= 3.22
Probability > F	= 0.043
R squared	= 0.315
Adjusted R square	= 0.217
Root MSE	= 6.741

Prob > F = 0.043.

If p value is less than level of significance we can reject null hypothesis.

At 10%, 0.04 < 0.10, reject null hypothesis

At 5%, 0.04 < 0.05, reject null hypothesis

At 1%, 0.04 > 0.01 accept null hypothesis

So normally we use 95% confidence, or 5% level of significance. So we reject null hypothesis. So we reject $\beta_0 = \beta_1 = \beta_2 = 0$. So we can say atleast one variable should be significant.

VARIABLE SELECTION - Automated Output as generated below.

total medal	Coeff	Std Error	t	P>|t|	[95% Conf.	Interval]
cen-lat	.522	.188	2.77	0.012	0.129	.915
elev	.003	.003	0.83	0.415	-0.004	.011
logpop	2.146	.996	2.15	0.043	0.073	4.219
-cons	-54.52	21.9	-2.48	0.022	-100.227	-8.827

-cons → Constant term

Normally in eqn, number of medal = $\beta_0 + \beta_1$ (latitude) + β_2 (elevation) + β_3 (logpop)

β_0 is number of medal when all other variables (Independent) is 0.

First Column is Coeff that is coefficient, so the equation become.

number of medals = $-54.52 + 0.522$ (latitude) $+ 0.003$ (elevation) $+ 2.146$ (log population)

Interpretation, For every additional degree of latitude, the expected number of medals increase by 0.523 on average, holding all other variable constant.

Suppose we want to estimate for INDIA: latitude = 52.2, elevation = 30.1m, Pop = 16,500,000

log pop = 16.62.

Number of medal = $-54.582 + 0.523(52.2) + 0.003(30.1) + 2.146(16.6) = 8.557$

So, India will win 8 medals

But in actual, INDIA got 24 medals, error (IND) = 24 - 8.6 = +15.4.

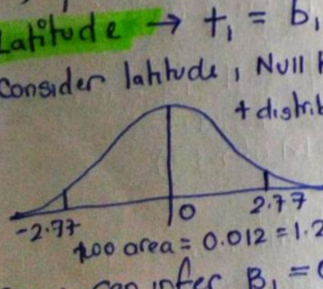
Std Error is standard error which is average error in the given sample.

t value is t statistic, higher the value of t, higher is the importance of the variable. So, if t value is +ve then it is +ve related and if t value is -ve then it is -ve related.

So, by t value we can say, latitude is most important then population then elevation

Latitude → $t_1 = b_1 / SE_1 = 0.522 / 0.189 = 2.77$

Consider latitude, Null hypothesis: $\beta_1 = 0$, $b_1 = 0.522$, $t_1 = 2.77$



If the null hypothesis is true ($\beta_1=0$), the chance of getting sample as extreme as we did is 1.2%.

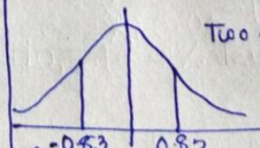
Two area = 0.012 = 1.2% (P)

So, we can infer $\beta_1 = 0$ is rejected because of too low (1.2%).

Elevation = 0.031/0.003 = 0.83

Null hypothesis, $\beta_2 = 0$, $b_2 = 0.0317$

$t_2 = 0.83$



Two area = 0.415

= 41.5%

(p value)

If null hypothesis is true ($\beta_2 = 0$), then 41.5% chance of getting it.

So elevation has no impact on total medal.